Log-Linear Analysis and Generalized Linear Models

Stephen Fisher

stephen.fisher@sociology.ox.ac.uk

Course website: http://users.ox.ac.uk/∼polf0050/page7.html

Outline

- Log-linear analysis
- Generalized Linear Models (GLMs)

Aims: To provide a basic introduction to log-linear analysis and the generalized linear models framework. The aim of the latter is to briefly illustrate some of the mathematical links between the models covered in the central four weeks of the course.

Log-linear modelling

This technique is used to analyse contingency tables, where the cells are frequencies. A two-way $I \times J$ table has the following form.

		Y			
		1	2	$\ldots$	J
X	1	m_{11}	m_{12}	$\ldots$	m_{1J}
	2	m_{21}	m_{22}	$\ldots$	m_{2J}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	I	m_{I1}	m_{I2}	$\ldots$	m_{IJ}

In log-linear analysis we use a Poisson (or similar) model the log of the counts as a linear function of some parameters.

The saturated log-linear model can be written as,

$$log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \Leftrightarrow m_{ij} = e^{\mu} e^{\lambda_i^X} e^{\lambda_j^Y} e^{\lambda_{ij}^{XY}} \tag{1}$$

where

- μ is a general mean (but not necessarily the mean)
- λ_i^X are row effects
- λ_j^Y are column effects
- λ_{ij}^{XY} are interaction terms

At the moment there are more parameters than there are cells in the table, so we need further identifying restrictions.

One set of restrictions, known as corner constraints, requires $\lambda_1^X = \lambda_1^Y = \lambda_{i1}^{XY} = \lambda_{1j}^{XY} = 0$ for all i and j.

For a 2×2 table the cell counts would therefore have the following form:

		Y	
		1	2
X	1	e^{μ}	$e^{\mu}e^{\lambda_2^Y}$
	2	$e^{\mu}e^{\lambda_2^X}$	$e^{\mu}e^{\lambda_2^X}e^{\lambda_2^Y}e^{\lambda_{22}^{XY}}$

So the odds ratio for the table is,

$$\frac{\frac{m_{11}}{m_{12}}}{\frac{m_{21}}{m_{22}}} = \frac{m_{11}m_{22}}{m_{12}m_{21}} = \frac{e^{\mu} \times e^{\mu}e^{\lambda_2^X}e^{\lambda_2^Y}e^{\lambda_{22}^{XY}}}{e^{\mu}e^{\lambda_2^X} \times e^{\mu}e^{\lambda_2^Y}} = e^{\lambda_{22}^{XY}} \tag{2}$$

So the interaction term λ_{22}^{XY} is a log odds ratio.

If the $\lambda_{22}^{XY} = 0$ there is no association and the model becomes the independence model,

$$log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y \tag{3}$$

Types of Independence

Loglinear models can incorporate more that two categorical variables and are useful for analysing patterns of association. Equally, we can use them to identify different types of independence.

Consider three categorical variables X,Y and Z.

- X,Y and Z are *mutually independent* if there is no association between any pair of variables. This is the case if the (X,Y,Z) model

  $$log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \tag{4}$$

  fits the data.

- X is *jointly independent* of Y and Z if there is no association between either X and Y or X and Z. This is the case if the (X,YZ) model

  $$log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ} \tag{5}$$

fits the data.

- X and Z are *conditionally independent given* Y if X and Z are independent within each level of Z. This is the case if the (XY,YZ) model

$$log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} \tag{6}$$

fits the data.

The concept of conditional independence is key to understanding Factor Analysis and Latent Class Analysis.

Death Penalty Example

		Death Penalty	
Defendant's Race	Victim's Race	Yes	No
White	White	19	132
	Black	0	9
Black	White	11	52
	Black	6	97

These data, from Radelet (1981) regarding homicide defendants in 20 Florida counties during 1976-1977, raise various questions, e.g.

- Is the victim's race associated with the defendant's race?
- Is the death penalty associated with the defendant's race?
- Is the death penalty associated with the victim's race?

- If there is an association between the death penalty and the defendant's race, does it depend on the victim's race? i.e. is there a three-way interaction?

While 12% of whites and 10% of blacks received the death penalty, when we break the rates down according to the victim's race we get:

	Victim's Race	
Defendant's Race	White	Black
White	14%	0%
Black	21%	6%

So while the *marginal* association between the death penalty and defendant's race is slightly to the benefit of black defendants, the *partial* association is clearly in favour of whites.

- This seems to be an example of Simpson's Paradox.

However, if we build a log-linear model, both the three-way interaction and the interaction between defendant's race and the use of the death penalty are statistically insignificant:

```
. xi: poisson count2 i.dr*i.vr i.dp*i.vr
i.dr              _Idr_1-2            (naturally coded; _Idr_1 omitted)
i.vr              _Ivr_1-2            (naturally coded; _Ivr_1 omitted)
i.dr*i.vr         _IdrXvr_#_#         (coded as above)
i.dp              _Idp_1-2            (naturally coded; _Idp_1 omitted)
i.dp*i.vr         _IdpXvr_#_#         (coded as above)
note: _Ivr_2 dropped due to collinearity

Poisson regression                                Number of obs   =          8
                                                  LR chi2(5)      =     379.22
                                                  Prob > chi2     =     0.0000
Log likelihood = -19.607184                       Pseudo R2       =     0.9063

------------------------------------------------------------------------------
      count2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Idr_2 |  -2.256065   .3169115    -7.12   0.000       -2.8772    -1.63493
      _Ivr_2 |  -.5668625   .1622843    -3.49   0.000     -.8849338   -.2487912
 _IdrXvr_2_2 |   3.112116   .3497906     8.90   0.000      2.426539    3.797693
      _Idp_2 |   -2.60269   .3664141    -7.10   0.000     -3.320848   -1.884531
 _IdpXvr_2_2 |   .8426789   .4133831     2.04   0.042      .0324629    1.652895
       _cons |   4.582501   .1008086    45.46   0.000       4.38492    4.780083
------------------------------------------------------------------------------

. poisgof

         Goodness-of-fit chi2  =  1.105725
         Prob > chi2(2)        =    0.5753
```

This analysis does not *prove* that defendant's race is conditionally independent of the death penalty given the victim's race, rather there is insufficient evidence in this data to suggest that there is an association between defendant's race and the death penalty given the victim's race.

- Other studies suggest there is such an association.

Relationship between Log-linear analysis and Logit

Consider the simple 2×2 situation with corner-constraints. The fitted values were

		Y	
		1	2
X	1	e^{μ}	$e^{\mu}e^{\lambda_2^Y}$
	2	$e^{\mu}e^{\lambda_2^X}$	$e^{\mu}e^{\lambda_2^X}e^{\lambda_2^Y}e^{\lambda_{22}^{XY}}$

Suppose that Y can be considered to be a dependent variable and we are interested in the odds of being at level 2 (rather than level 1). Also, let $x = 1$ if X is at level 2, and 0 otherwise.

Then we could fit the following logit model,

$$\log\left(\frac{\Pr(Y=2)}{\Pr(Y=1)}\right) = \lambda_2^Y + \lambda_{22}^{XY} x \tag{7}$$

As for the standard logit model

- the intercept, λ_2^Y is the baseline log odds, i.e. the odds that $Y = 2$ given that $X = 1$.

- The coefficient of x is the difference between the log odds that $Y = 2$ between levels 1 and 2 of X, i.e. it is the log odds ratio.

So if one of your categorical variables is clearly a binary dependent variable, logisitic regression is a more natural, but statistically equivalent, alternative to loglinear modelling.

Similarly if one of your categorical variables is a multi-category dependent variable, then multinomial logit may be a useful, but statistically equivalent, alternative to loglinear modelling.

Extensions and Software

Loglinear analysis has been extended in two main ways:

- **Graphical models** which map the patterns of association between many categorical variables

- Models for **square tables** or matched-pairs, especially mobility tables. Some of these are actually non-linear, but the framework and notation for log-linear models tend to be a basic component of the models.

Simple loglinear models of the kind discussed here can be estimated in Stata with `poisson` or more easily with the `ipf` or `loglin` plug-ins. SPSS and the `loglin` package in R are also possibilities.

For more complicated loglinear and nonlinear models for categorical variables there is *Lem* (free) and *Latent GOLD* (costly) by Jeroen Vermunt, or various R packages, especially David Firth's Generalized Nonlinear Models `gnm` package.

Generalized Linear Models

Most of the models covered so far in the course have been examples of the broader class of *generalized linear models*.

The generalized linear model has three components:

1. A *random component*: the dependent variable Y_i which, conditional on the independent variables, follows one of the distributions in the exponential family including: normal, Poisson, binomial, gamma, or inverse-Gaussian.

2. A *linear predictor* $\eta_i = \mathbf{x_i}\boldsymbol{\beta}$ on which the dependent variable depends

3. A *link function* $L(.)$ that transforms the expectation of the dependent variable $\mu_i \equiv E(Y_i)$ to the linear predictor η_i. Common link functions include:
   - the identity link: $L(\mu_i) = \mu_i$
   - the log link: $L(\mu_i) = log(\mu_i)$
   - the logit link: $L(\pi_i) = log(\frac{\pi_i}{(1-\pi_i)})$
   - the probit link: $L(\pi_i) = \Phi(\pi_i)$, where $\Phi(.)$ is the cdf of the unit-normal distribution

Further Reading

Many textbooks include more thorough treatment of the material in this lecture. One that I would particularly recommend is:

Agresti, Alan: *Categorical Data Analysis*, (Wiley).

This is good for loglinear analysis for which he also has a 1990 book, but also other models for mobility tables.

Perhaps the simplest outline of GLMs is Jeff Gill's small green Sage book on the subject.